# Getting down to details

**MicroRNAs that tweak gene expression, single nucleotide polymorphisms in population genetics, and individual genome sequencing: Caitlin Smith takes a look at three fast-moving areas in genomics.**

Over the past few years, genomics researchers have been getting to grips with a 'new' genome element — microRNA (miRNA). Although a small number of miRNAs have been familiar to developmental biologists for years, a plethora of miRNAs has recently been discovered in animal and plant genomes. More than 200 miRNAs have been identified in mammalian genomes, but their functions mostly remain a mystery. Silencing gene expression in a similar way to small interfering RNAs (see *Nature* 431, 350; 2004), mammalian miRNAs are implicated in the control of cell and tissue differentiation, apoptosis, insulin secretion, fat metabolism and cancer.

"We are now aware that there is substantially more transcription from human chromosomes than can be accounted for by the current predictions of human genes," says Frank Slack at Yale University, New Haven, Connecticut. Slack is studying the apparent involvement of the miRNA *let-7* in lung cancer and the implications of its ability to suppress translation of the oncogene *RAS*. "Many miRNAs are mapping to disease loci where previously a gene was not found," he says.

miRNA research is a typical microcosm of the variety of disciplines and techniques that are required to make sense of the genome — computational biology, bioinformatics and comparative genomics to predict candidate miRNAs, followed by classic 'wet biology' to validate the candidates and study their expression and function. And as more labs are gearing up to study miRNAs, commercial products tailored to help them are coming onto the market.

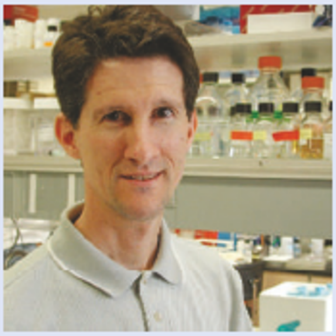
**Micro solution: Ambion's flashPage isolates small RNAs.**

The technical problems of detecting miRNAs in total cellular RNA stem from their small size and often low abundance. Produced from a larger precursor molecule, mature miRNAs are RNA hairpins of 17–23 nucleotides, which bind to complementary sequences in their target messenger RNAs (mRNAs) and prevent translation. The general techniques for detecting and isolating miRNAs from cellular RNA are those used for other small RNAs. A first step could be spin-column fractionation of RNA to remove larger RNAs, using columns such as the Amicon YM-100 from Millipore of Bedford, Massachusetts, which will remove RNAs of more than 75 bases, or the PureLink miRNA isolation kit from Invitrogen of Carlsbad, California, with a 200-nucleotide limit. Qiagen of Valencia, California, has a small RNA protocol for their widely used RNeasy system, which will remove RNAs of more than 200 bases from total cellular RNA. To get even closer to mature miRNA length, RNA specialists Ambion of Austin, Texas, sells a flashPage, a gel-based fractionation machine for the rapid isolation of small nucleic acids of around 40 bases.

After initial preparation, specific miRNAs

## BIG TASKS FOR SMALL MOLECULES

Having helped to identify many miRNAs, Christopher Burge and his colleagues at the Massachusetts Institute of Technology are one of many teams now tackling an even bigger job — to find out which genes are regulated by the known miRNAs, and how they fit into physiological pathways.



**James Carrington: taking a systems look at miRNA.**

Finding targets begins computationally, using the TargetScan algorithms developed by Benjamin Lewis working with Burge and with David Bartel at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. These algorithms "rely on evolutionary conservation of segments complementary to the microRNA 'seed' region in the 3' untranslated regions of orthologous genes from multiple vertebrate organisms", says Burge. The seed region, six or seven bases at the 5' end of the miRNA, is thought to be key to specifying which genes an miRNA will regulate. Targets have been verified in Bartel's lab using a dual luciferase reporter system, which measures the effect of predicted miRNA interaction sites on protein production in cultured human cells. In a computational analysis published earlier this year, Lewis, Burge and Bartel estimated that more than a third of our genes might be regulated by miRNAs.

The task will be complicated by the fact that an miRNA may regulate as many as 200 genes, according to a computational study by Nikolaus Rajewsky and his colleagues at New York University and Rockefeller University, using their PicTar algorithm to identify miRNA targets. Other software for miRNA target prediction includes miRANDA from Anton Enright and his colleagues at the Memorial Sloan-Kettering Cancer Center in New York and DIANA-microT from Artemis Hatzigorgeou and Axel Bernal at the University of Pennsylvania, Philadelphia.

Frank Slack's team at Yale University uses *in situ* hybridization, northern blots and fluorescent protein fusions to find when and where miRNAs and their targets are expressed. "We use genetics and RNA interference to reduce the expression of potential targets to see if we suppress the effects of a mutation in the corresponding miRNA, and use reporter gene assays to test if the miRNA-complementary sites function in gene regulation," he says.

"The classic tools of developmental biology and physiology are needed to correlate miRNA expression and targeting to biological function," agrees James Carrington at Oregon State University, Corvallis, who is looking at pathways regulated by miRNAs in *Arabidopsis*. "miRNA sensors involving miRNA target sites within gene constructs expressing a fluorescent protein are quite useful in understanding spatial and temporal miRNA expression and activity patterns," he says. But to address the question of how miRNAs integrate with cellular pathways, "the more quantitative approaches using the tools of systems biology and computational analysis are the trend in this lab", he says. C.S.

in the sample can be detected by techniques such as northern analysis, PCR and microarrays. But how do you know what you're looking for? Much of the groundwork in miRNA identification has been laid by large-scale genomics projects that used computational techniques to predict miRNA genes followed by cloning and validation of the predicted sequence. The public miRNA registry currently holds around 1,650 entries for published predicted miRNAs. The big projects now under way are to determine which genes the miRNAs are targeting (see 'Big tasks for small molecules', page 991).

Northern analysis is still the standard for detecting and quantifying miRNA expression. "Northern blotting, even if time consuming, is by far the best technique to study miRNA expression because of its sensitivity and quantitativity," says Jiahuai Han at the Scripps Research Institute, La Jolla, California, who is looking at the mechanisms by which miRNAs affect the stability of their target mRNAs. "Primer extension has the advantage of being quicker but, unfortunately, is less quantitative," he says.

Integrated DNA Technologies (IDT) of Coralville, Iowa, sells miRNA tools to increase the sensitivity of northern analysis. Its StarFire kit for probe labelling makes labels composed of 10 $^{32}$P-alpha-dATPs rather than the more usual single $^{32}$P-gamma-ATP. "We use a special template and reaction conditions that give a 10-base tail with almost no heterogeneity," explains Mark Behlke, vice president of molecular genetics at IDT, "so all probe molecules are the same — it is very different from other tailing procedures." Manufacturers are also gearing up to make miRNA-specific probes; miRCURY LNA (locked nucleic acid) detection probes for all known miRNAs are available from Exiqon of Vedbaek, Denmark, for example, and can be used for *in situ* hybridization, northern analysis, PCR and gene knockdown.

Another choice for miRNA detection is Ambion's mirVana miRNA detection kit. Ambion claims that its assay is 100–500 times more sensitive than northern analysis, as the radiolabelled probes are hybridized in solution instead of on a membrane as in northern blotting. The company claims that this method gives the researcher a better shot at detecting and quantifying low-abundance miRNAs because the probe and target have more opportunities to bind when in solution.

Taking the PCR road, Applied Biosystems of Foster City, California, is soon to launch a new TaqMan microRNA assay for miRNA detection and quantitation, which the company claims will detect only mature miRNAs and not precursors. According to Marcum Bell, product manager of gene-expression assays at Applied Biosystems, the assay "uses specific stem-looped primers for reverse transcription of the mature miRNA, followed by quantitative real-time PCR." A claimed advantage of the new assay is its wide dynamic range of up to 7 log units, enabling detection of both low- and high-abundance miRNAs.

For an alternative to PCR-based miRNA assays, US Genomics of Woburn, Massachusetts, recently unveiled its Trilogy 2020 Single Molecule Analyzer for the high-throughput detection and quantitation of single molecules of nucleic acid without amplification. The Trilogy 2020 can be used along with the company's Direct miRNA Assays for miRNA work. The assay includes two fluorescently tagged probes (tags can be red, blue or green) that are designed to hybridize to the miRNA of interest. Specificity relies on the very high likelihood that only the target miRNA will hybridize to both probes. After hybridization,

US GENOMICS

**The Trilogy Single Molecule Analyzer can be used for miRNA detection.**

## GENOTYPING GETS UP TO SPEED

**At the high-throughput end of multiplex SNP genotyping, Illumina of San Diego, California, is currently beta testing the Sentrix Human-1 BeadChip, containing more than 100,000 SNPs, nearly 30,000 of which are located in genes, with another 40,000 within 10 kb of genes. The company is developing BeadChips containing 250,000 and 500,000 SNPs for release next year, which will make it possible to genotype 1 million SNPs on just a pair of chips.**

**Using a different approach to SNP genotyping, the LightTyper Genotyping System from Roche Applied Science of Indianapolis, Indiana, is designed for the heavy-duty end of the market, where thousands of samples may have to be genotyped each day. After PCR amplification of genomic samples in 96- or 384-well plates using a standard thermal cycler, plates are transferred directly to the LightTyper and genotyped within 10 minutes, using the melting points of fluorescently labelled probes hybridized with the SNPs as the detection system. Probe-target complexes with different melting points reflect the presence of different alleles, and show up as allele-specific peaks in the melting curves. Because many samples can be tested simultaneously, "the LightTyper instrument is mainly used for SNP genotyping, in particular for disease association studies," says Burkhard Ziebolz of science communications at Roche Diagnostics in Mannheim, Germany.**

**The Luminex xMAP platform for multiplex genotyping is used by several genetic diagnostics service companies, including TmBioscience of Toronto, Ontario, which has developed the first Food and Drug Administration-approved multiplexed test for cystic fibrosis mutations, and Tepnel LifeCodes of Manchester, UK, whose speciality is HLA DNA typing.**

**Roche's LightTyper speeds up high-throughput genotyping.**

TOM MERCER/CLEVELAND CLINIC

**For less-intensive SNP detection, the READIT SNP genotyping system from Promega of Madison, Wisconsin, can be scaled up or down. It uses the company's READase-mediated destabilization of perfectly matched probe-target complexes coupled with a luciferase reporter assay for the ATP generated. With appropriately designed probes, the system can detect SNPs, insertions, deletions and chromosomal translocation, and can estimate allele frequency and carry out allele-correlation studies. And PerkinElmer of Boston, Massachusetts, have SNP detection kits in their established AcycloPrime range.** C.S.

the sample is moved by microfluidics through a glass capillary, where lasers excite the probes at different wavelengths. A target miRNA molecule is counted when photons of both colours are emitted simultaneously.

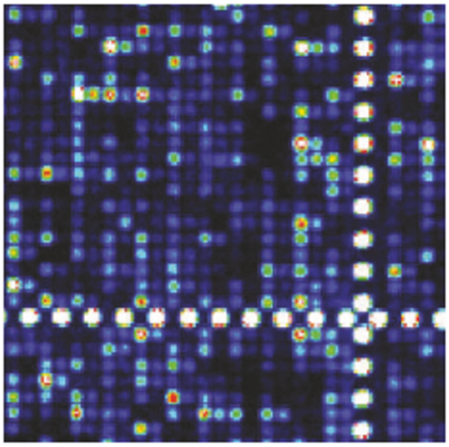
Both conventional microarrays and bead-based multiplex assay platforms such as xMAP from Luminex of Austin, Texas, can be used to study miRNA expression, and a number of companies offer miRNA products designed for use with microarray systems. PerkinElmer of Boston, Massachusetts, sells a MICROMAX ASAP labeling kit for miRNAs for detection by the tyramide signal amplification (TSA) method, while the Array 900miRNA labeling kits from Genisphere of Hatfield, Pennsylania, are designed to label miRNAs and other small RNAs with Genisphere's 3DNA dendrimers. If you don't want to do it yourself, companies such as molecular diagnostics specialists Genaco of Huntsville, Alabama, and genetic services company DNAVision of Charleroi, Belgium, offer miRNA expression profiling and quantitation using Luminex xMAP technology. LC Sciences of Houston, Texas and Icoria of Research Triangle Park offer microarray-based miRNA detection covering all miRNAs currently listed in the public miRNA registry.

## Differences matter

If miRNAs are the new kid on the block in genomics, single nucleotide polymorphisms (SNPs) are already big business (see 'Genotyping gets up to speed', opposite). Your DNA is 99.9% identical to that of another unrelated human, but it is that last 0.1% that interests researchers. Much of the difference is made up of SNPs, which are sites in DNA that differ by a single base. Groups of SNPs close to one another on a chromosome are called blocks, and are usually inherited together as a haplotype, thus providing a convenient marker for the other genes in the block. The HapMap project, run by the International HapMap Consortium, aims to create a map of these haplotypes and their SNP tags for future research (see 'SNPs and human disease', below).

Using SNP tags, scientists can more efficiently scan an individual's genome for association with phenotypes, such as disease susceptibility, or reactions to drugs or vaccines. Launched in October 2002, the HapMap project hoped to complete the mapping of one million SNP markers by September 2005. When it achieved this goal months ahead of schedule, the consortium announced this February that it will step up its efforts in the second phase to create an improved map that is five times denser than the first draft. This will enable geneticists to zero in on smaller areas of the genome, locating targets more precisely by using more SNP signposts, increasing coverage from one SNP every 3,000 bases (at present) to one every 600 bases.

Vital to phase 2 is Perlegen Sciences of Mountain View, California, which is testing 4.6 million SNPs from public databases for addition to the HapMap. Last September, funded by a grant from the US National Human Genome Research Institute, Perlegen began using high-density oligonucleotide array technology from Affymetrix of Santa Clara, California, to genotype more than 2.25 million unique SNPs from the four HapMap study populations. Perlegen's original goal was to catalogue 600 million genotypes; the new funding in phase 2 should result in more than a billion. The human genome is thought to contain about 10 million SNPs, but not all of these will be useful predictors of disease.

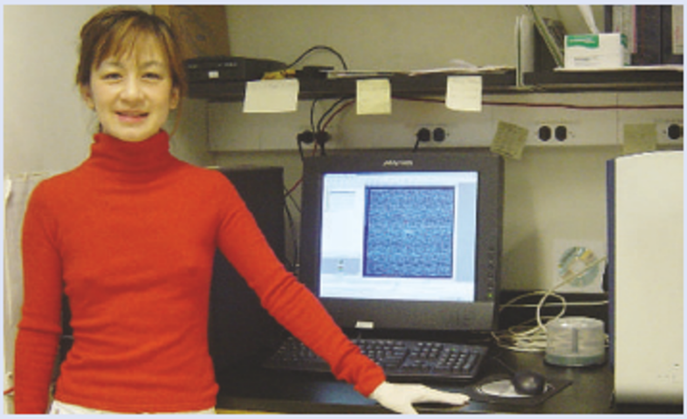


**Patterns of life: DNA bound to a small region of an Affymetrix 100K GeneChip Set.**

David Cox and his colleagues at Perlegen aim to narrow the field. They have analysed the most common SNPs by mapping 1.5 million SNPs for 71 people from three different ethnic groups: European American, African American and Han Chinese American. The aim is to obtain a high-quality subset of SNPs

## SNPs AND HUMAN DISEASE

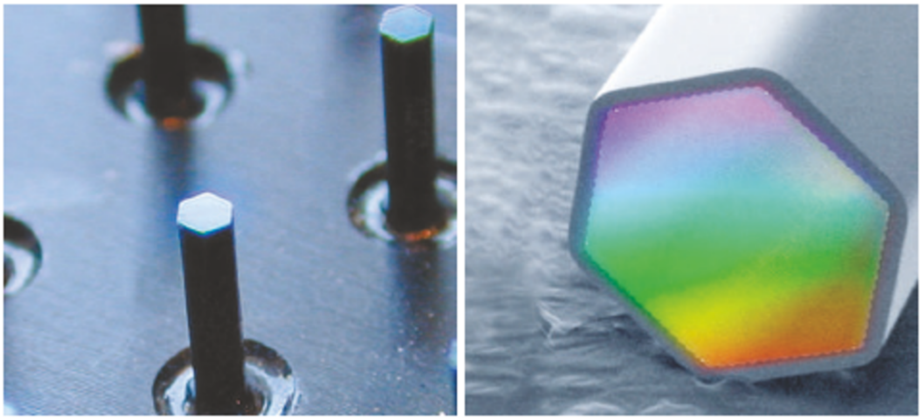
One goal of the HapMap project is to help reserachers find SNPs associated with human disease. Josephine Hoh at Yale University's School of Public Health and colleagues at Rockefeller University, New York, and the National Eye Institute in Bethesda, have identified an SNP associated with age-related macular degeneration (AMD), a major cause of blindness in people over 60. The SNP is in the gene for complement factor H, leading to a tyrosine to histidine mutation. The researchers studied 96 patients with AMD and 50 healthy controls, and measured the frequency of over 116,000 SNPs in each group. "For the initial screen, we used Affymetrix's set of 100K SNP chips," says team member Robert Klein. "To identify the putative causal mutation, we used PCR to amplify each exon in a number of samples and then resequenced to find all variants in the exons."



**Josephine Hoh uses SNP arrays to find mutations associated with disease.**

Ho and her colleagues found that caucasian patients with AMD are at least four times more likely than usual to have this SNP. How the change causes AMD is not yet known, and one of the next directions for her lab "is to figure out the functional mechanism of complement factor H in the pathogenesis of AMD", says Hoh. There are a few clues. The amino-acid change lies in a part of factor H that interacts with C-reactive protein and heparin, both known to be associated with AMD. And factor H is known to regulate components of the immune system that are found in drusen, fatty deposits that accumulate in the macula with age. In people with AMD, the drusen are larger and more numerous, killing cells needed to nourish adjacent retinal photoreceptors, which eventually results in loss of sight.

SNP mapping is also underway in animal models of human disease. Kent Hunter at the National Cancer Institute (NCI) in Bethesda, Maryland, uses SNPs to look for cancer-modifying genes in mice. "Ultimately, we hope to identify the particular polymorphic gene or genes that modulate metastatic efficiency," he says. Maxwell Lee at the NCI is interested in how genetic variation determines gene expression and phenotypes in human cancer and uses SNPs to search for epigenetic markers. "We need to understand more dynamic aspects of the genome including interactions between SNPs and other downstream targets such as chromatin, DNA methylation and gene expression," he says. C.S.



**A Sentrix array matrix (left) used for genotpying and one of its fiber bundles (right).**

ILLUMINA

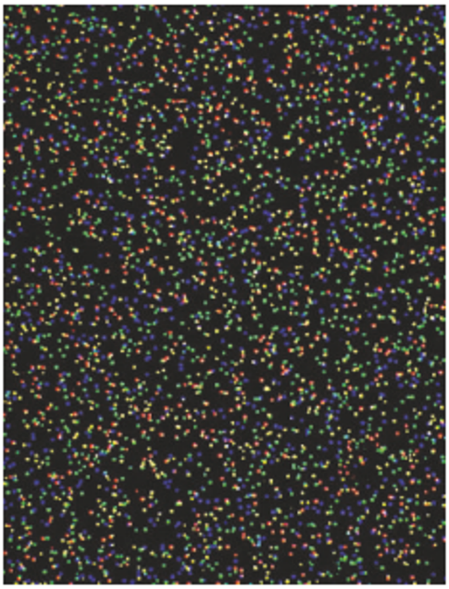
for disease prediction, and to make these subsets more useful by learning more about the frequencies of alleles and how they are correlated with one another.

## The $1,000 genome?

As Illumina closes in on the million-SNP assay (see 'Genotyping gets up to speed', page 992), others are striving for the $1,000 genome — a quick and cheap method of sequencing individual genomes. This somewhat arbitrary goal has caught the fancy of scientists and is being competitively pursued by companies — fuelled in part by a $500,000 cash prize offered by the J. Craig Ventner Science Foundation to whoever gets there first.

For this goal to become reality, a new method must replace the much-loved Sanger method and its offspring. The leading alternative is single-molecule-based sequencing, also known as sequencing by synthesis. VisiGen Biotechnologies in Houston, Texas, uses this method with single-pair fluorescence resonance energy transfer (spFRET) as the detection technology. The donor fluorophore is attached to a DNA polymerase that sits on the template, while a colour-coded acceptor fluorophore is attached to the gamma-phosphate of a nucleoside triphosphate. When the nucleotide is incorporated into DNA, the donor fluorophore stimulates the acceptor to emit a characteristic fluorescent signal (measured as emission wavelength and intensity) that indicates its base identity — each base is a different colour.

"The donor fluorophore acts as a punctuation mark between nucleotide incorporation events," explains Susan Hardin, president and chief executive of VisiGen. Massively parallel arrays of these reactions produce a high-throughput sequencing system without the need for electrophoresis, cloning or PCR. Hardin points out that using a donor label on an immobilized polymerase "minimizes background, increases consistency of the signal during the extension, and increases read length." The advantage of labeling the nucleotide on its terminal gamma phosphate is that the fluorophore does not become part of the nascent DNA strand.



**Light fantastic: the first cycle in a round of sequencing by Solexa's method.**

SOLEXA LTD

The sequencing by synthesis method of Solexa, based in Little Chesterford, UK, (which recently merged with Lynx Therapeutics of Hayward, California) differs from that of VisiGen mainly in the detection method. In the Solexa technology, the different types of fluorescently labelled nucleotide incorporated into the DNA strand are detected by excitation with an external light source. The cycle of nucleotide incorporation, detection and identification is repeated about 25 times to read the first 25 bases in each oligonucleotide in an array of millions of single-stranded genomic DNA fragments.

According to Simon Bennett, Solexa's business development director, one advantage of the company's system is the small sample volume required — only a few picograms. "Biobanks may need to seriously reconsider how to collect and store samples, and may wish to explore cheaper options," says Bennett. "With the emerging technologies the cost of analysing samples, and how much sample is needed for each subject, is almost certain to reduce dramatically." He also remarks that the short sequence read lengths of 25–30 nucleotides in Solexa's method may allow a degraded sample to be resequenced, rendering a previously useless sample once again viable.

A stab at the $1,000 genome is also being taken by 454 Life Sciences, a subsidiary of Curagen in Branford, Connecticut. The 454 whole-genome sequencing system, a prototype of which was recently installed at the Broad Institute in Cambridge, Massachusetts, uses the patented light-emitting 'pyrosequencing' chemistry developed by Pyrosequencing of Uppsala, Sweden, (now renamed Biotage after its recent takeover of that company) and microfluidics nanotechnology developed by 454. The pyrosequencing technique was exclusively licensed to 454 in 2003 for the purposes of developing it for whole-genome sequencing. Biotage retains the rights to use the technology in its gene-analysis products, such as the PyroMark range of genetic tests launched earlier this year. Last month, Roche Applied Science signed an exclusive licence with 454 to develop and sell the 454 whole-genome sequencer.

So when can we expect the first $1,000 genome? "2010 to 2012," says Hardin. Bennett is less definite, but thinks that Solexa will be offering a $1,000 genome product before 2016, and "probably before the end of this decade."

## Future prospects

Looking to what's next on the genomics agenda, Carrington settles for understanding the mechanisms of genome evolution and adaptation, especially the evolution of new functions. He cites the recent discovery of possible "multigenerational sequence caches to restore lost information from a genome," in which plants have been shown to inherit DNA sequences not apparently present in parental genomes yet found in previous generations; one explanation might be caches of RNA. "Understanding the mechanisms of formation and adaptation of new miRNA genes, and more broadly, the derivation and deployment of small RNA-based regulation at the transcriptional and post- transcriptional levels, represent a major set of questions under the umbrella of genome evolution," he adds.

On the biomedical front, Hardin points out that geneticists will soon be facing serious ethical considerations. "Who should have access to an individual's genome sequence?" she asks. "What should one be told about (potential) defects in one's genome sequence? When? How can we best safeguard the privacy of genome-sequence information?" Given the rapid advance of genomics, scientists will be having to answer these questions sooner than they might have thought. ■

**Caitlin Smith is a science writer based in Portland, Oregon.**

miRNA Registry
▸ **www.sanger.ac.uk/Software/Rfam/mirna/index.shtml**
International HapMap Consortium
▸ **www.hapmap.org**